

## SPECTRAL REGRESSION DISCRIMINANT ANALYSIS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Yinsong Pan, Junyuan Wu, Hong Huang, Jiamin Liu

Key Lab. on Opto-electronic Technique and systems, Ministry of Education, Chongqing University, Chongqing, P.R.  
China, 400044

ICWG III/VII

**KEY WORDS:** Dimensionality reduction, Hyperspectral image classification, manifold learning, Eigen-decomposition, Spectral regression discriminant analysis, Embedding function

### ABSTRACT:

Dimensionality reduction algorithms, which aim to select a small set of efficient and discriminant features, have attracted great attention for Hyperspectral Image Classification. The manifold learning methods are popular for dimensionality reduction, such as Locally Linear Embedding, Isomap, and Laplacian Eigenmap. However, a disadvantage of many manifold learning methods is that their computations usually involve eigen-decomposition of dense matrices which is expensive in both time and memory. In this paper, we introduce a new dimensionality reduction method, called Spectral Regression Discriminant Analysis (SRDA). SRDA casts the problem of learning an embedding function into a regression framework, which avoids eigen-decomposition of dense matrices. Also, with the regression based framework, different kinds of regularizers can be naturally incorporated into our algorithm which makes it more flexible. It can make efficient use of data points to discover the intrinsic discriminant structure in the data. Experimental results on Washington DC Mall and AVIRIS Indian Pines hyperspectral data sets demonstrate the effectiveness of the proposed method.

### 1. INTRODUCTION

Hyperspectral image processing has become one of the hot topics in recent twenty years, because the hyperspectral images contain much useful information, and the information plays an important role in the development of society and economy. Hyperspectral images combine imaging techniques with the spectral techniques together, are widely used in civil and military areas. Compared with the moderate dimension images, hyperspectral images can get a finer spectral curve, also provide stronger evidence to the recognition of ground covers. Though the high resolution, the hyperspectral images inevitably contribute to information redundancy and difficulty to data processing. In order to solve these problems, we should reduce the dimensions of the images, because with the increase in the number of the dimensions, the number of the training samples must be increased exponentially [1].

Broadly speaking, there are many approaches to the dimensionality reduction [2], such as Principal Component Analysis (PCA) [3], Linear Discriminant Analysis (LDA) [4], Kernel Principal Component Analysis (KPCA) [5] and Neighborhood Preserving Embedding (NPE) [6]. These methods will result in a large consumption of time, money and energy, because they are all involved with the eigen-decomposition of dense matrices. As a matter of fact, the Principal Component Analysis (PCA) is a linear feature extraction

method, whose shortcoming is that it extracts the principal information, but ignores the secondary information [7]. The Kernel Principal Component Analysis (KPCA) will improve the time complexity of the eigen-decomposition of a kernel matrix and decrease the speed of feature extraction for test samples. The Linear Discriminant Analysis (LDA) cannot directly apply the corresponding discriminant criterion to case of that the within-class scatter matrix is singular. Since the main goal of the Neighborhood Preserving Embedding (NPE) is to preserve localities or similarity ranking, these algorithms is appropriate for retrieval or clustering rather than for classification.

To overcome the above shortcomings, a novel dimensionality reduction method, Spectral Regression Discriminant Analysis (SRDA), is presented in this paper. Spectral Regression Discriminant Analysis (SRDA) has recently emerged as a powerful and efficient tool for dimensionality reduction and manifold learning. This method uses information contained in the eigenvectors of a data affinity matrix to reveal the low dimensional structure in high dimensional data. The most popular manifold learning algorithms include Locally Linear Embedding [8], Isomap [9], and Laplacian Eigenmap [2]. However, these algorithms only provide the embedding results of training samples. With regression as the building block, various kinds of regularization techniques can be easily incorporated in SRDA which

makes it more flexible. And SRDA can be used in supervised, unsupervised and semi-supervised situations. It can make efficient use of both labeled and unlabeled points to discover the intrinsic discriminant structure in the data.

The rest of the paper is arranged as follows. Part 2 is the detailed description of our proposed method. Part 3 will discuss some results of related experiments and the last part, the 4th part will draw some conclusions about all mentioned.

## 2. SPECTRAL REGRESSION DISCRIMINATE ANALYSIS (SRDA)

The Spectral Regression Discriminant Analysis (SRDA) will be introduced by three subsequent aspects. First of all, a graph embedding view of dimensionality reduction is mentioned, which is the fundamental of the SRDA. Secondly, the Spectral Regression Discriminant Analysis (SRDA) is discussed. Last but not the least, theoretical analysis is described. Details will be given by follows.

### 2.1 Graph Embedding View of Dimension Reduction

Given a data sample  $\{x_i\}_{i=1}^m$  in  $R^n$ , the purpose of dimensionality reduction is to find a low dimension representation of  $\{x_i\}$ , here we name the low dimension representation  $\{z_i\}_{i=1}^m \subset R^d, d \ll n$ . In recent ten years, no matter what presentations, many methods have been contributed to deal with this problem. All related algorithms, such as PCA, NPE, LDA, KPCA, can be efficiently interpreted in a general graph embedding framework [5, 10, and 11].

Consider a graph  $G$  with  $m$  vertices, and a vector presents a data point.  $W$  is a  $m \times m$  matrix also is symmetrical and  $W_{ij}$  has the weight of edge joining points  $i$  and  $j$  [12]. Graph embedding is to present each vector by a low dimension vector.

Define  $y=[y_1, y_2, y_3, \dots, y_m]$  as the map from the graph to real line, let  $L = D - W$ ,  $D$  is a diagonal matrix with entries are the column sum of  $W$ , so the optimal  $y$  will be obtained through minimizing

$$\sum_{i,j} (y_i - y_j)^2 W_{ij} \quad (1)$$

under appropriate constraint. When the vertices  $i$  and  $j$  are mapped far apart, the objective functions incur a heavy penalty. Therefore, minimizing it is an attempt to ensure that if vertices  $i$  and  $j$  are “close” then  $y_i$  and  $y_j$  are close as well. So we can elaborate it in simple formulation as follows:

$$\sum_{i,j} (y_i - y_j)^2 W_{ij} = 2y^T L y \quad (2)$$

So, the minimum problem changes to find

$$y^* = \arg \min_{y^T D y = 1} y^T W y = \arg \min_y \frac{y^T W y}{y^T D y} \quad (3)$$

Also, we can obtain the optimal  $y$  by solve the maximum eigenvector of

$$W y = \lambda D y \quad (4)$$

Now define  $y_i = f(x_i) = a^T x_i$  as the map from the graph to real line, Eq.(1) can be written by:

$$a^* = \arg \min_a \frac{y^T W y}{y^T D y} = \arg \min_a \frac{a^T X W X^T a}{a^T X D X^T a} \quad (5)$$

There  $X=[x_1, x_2, \dots, x_m]$ . Correspondingly, the optimal  $a$  is the eigenvector of the minimum eigenvalue of the follow eigen-problem:

$$X W X^T a = \lambda X D X^T a \quad (6)$$

This method is named as Linear extension of Graph Embedding (LGE), and if we choose the different  $W$ , we get other's methods of dimensionality reduction, such as LDA, NPE and LPP. Generally speaking, solving the eigen-problem of the Eq.(4) calls for not only time but also memory. And when the number of features is larger than the sample data, there is no efficient approach. Though we can seek Singular Value Decomposition for help, the complexity of the computation will increase.

### 2.2 Spectral Regression Discriminant Analysis

Seeking new method to solve the problem in Eq.(6) is an urgent task. First of all, we introduce a theorem as follows:

**Theorem 1:** Define  $y$  as the eigenvector of the eigen-problem in Eq.(4), and whose eigenvalue is  $\lambda$ . If  $X^T a = y$ , so  $a$  is the eigenvector of the eigen-problem in Eq.(6) with the same eigenvalue  $\lambda$  [13].

From theorem 1, we can solve the eigen-problem in Eq.(6), and the linear embedding functions can be obtained by two main steps: firstly, get  $y$  by solve the eigen-problem in Eq.(4). Secondly, search for the optimal  $a$ , make it satisfy  $X^T a = y$ . As a matter of fact, the  $a$  maybe not exist. Practical solutions can be found through the follow equation:

$$a = \arg \min_a \sum_{i=1}^m (a^T x_i - y_i)^2 \quad (7)$$

Here  $y_i$  represent the  $i$ -th element of  $y$ .

The obvious advantage of the above two-way method is that the matrix  $D$  can be firmly proven to be positive definite and the solution to the eigen-problem in Eq.(4) is stable. In addition, the least square problem will be efficient solved since the related technique is nearly mature no matter the scale of the matrix is large or small [14].

When the number of the sample data is smaller than the features', the minimum problem shows ill posed. There will be a lot of solutions to the linear equation system  $X^T a = y$ . Common method is to bring in a norm

$$a = \arg \min_a \left( \sum_{i=1}^m (a^T x_i - y_i)^2 + \alpha \|a\|^2 \right) \quad (8)$$

This is called regularization, and the  $\alpha$  is the parameter of scale control [15]. The third advantage of two-way method is that with the regularization can be incorporated to the model of regression; we can obtain stable and significant solutions [15].

This algorithm conducts regression processing after spectral analysis of the graph, so we name it the Spectral Regression Discriminant Analysis, or SRDA in short form [16].

### 2.3 Theoretical Analysis

With the regularized least squares, the SRDA can get the embedding function. When the parameter to control the amount of shrinkage  $\alpha > 0$ , the regularized solution cannot fit the linear equation system  $X^T a = y$ , also  $a$  is not the eigenvector which satisfies the eigen-problem in Eq.(6) [17]. In addition, we should be care of that when the SRDA gives the exact solutions to the eigen-problem in Eq.(6). Here, we have the theorem follows:

**Theorem 2:** Suppose  $y$  is in the space spanned by row vector of  $X$ , the corresponding projective function  $a$  calculated in SRDA will be the eigenvector of the eigen-problem in Eq.(8) as  $\alpha$  decreases to zero.

In a broad sense, we can get following corollary:

**Corollary 1:** If the vectors of sample data are linearly independent, that is,  $\text{rank}(X) = m$ , all the projective functions in SRDA are the eigenvectors of the eigen-problem in Eq.(4) as  $\alpha$  decreases to zero. These solutions are identical to the linear graph embedding solutions in section 2.1 [18].

SRDA seeks the projective functions by regularized least square, and this is the necessary steps in supervised case as well as the unsupervised case. In supervised case, SRDA has linear-time complexity, which is only concerned with  $m$  and  $n$ , while LDA has cubic-time complexity, and it is related to the minimum of  $m$  and  $n$ . Obviously, SRDA shows more advantages than LDA. In unsupervised case, SRDA uses regression to find out the projective functions, and the time complexity can be computed linearly, in addition this process almost cost little memory. Some linear extension methods, such as LDA, NPE obtain the projective functions through solving the dense eigen-problem. They need cubic-time complexity cost with about  $(m+n) \cdot \min(m,n)$  cost of memory. Also, SRDA is superior to the mentioned approaches above.

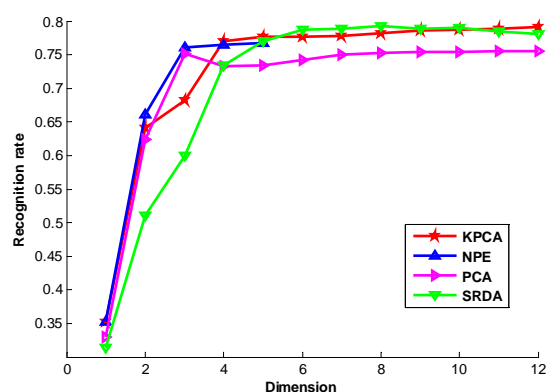
## 3. EXPERIMENTAL RESULT

In this part, we will carry out experiments on two dataset by the SRDA algorithm for classification to compare the effectiveness of SRDA with other methods. The first dataset is called Washington DC Mall, which has a dataset of 11414 samples which are all the labeled experiments, and the label ranges from 1 to 9. Each sample contains 12 bands. The second dataset is called AVIRIS Indian Pines, which has a dataset of 145\*145 samples, also the samples are all labeled, which ranges from 1 to 15. Each sample contains 220 bands.

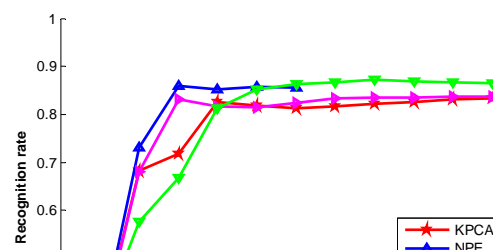
### 3.1 Experiments on Washington DC Mall

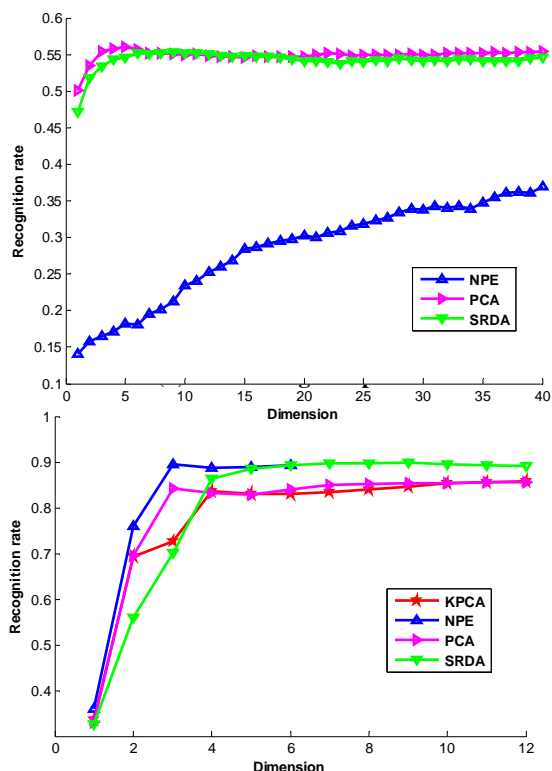
In order to investigate the performance of our proposed SRDA, we should start our experiments in following steps.

- 1) For each label, we find out 320 samples as experimental samples, that is to say, this dataset needs 320\*9 samples;
- 2) For the chosen 320 samples,  $l(=5, 10, 15, 20)$  datum are randomly selected for training and the rest 300 datum are used for testing;
- 3) The 1-nearest neighbor classifier is applied in PCA, KPCA, NPE and SRDA subspace. For PCA, KPCA and SRDA, the number of the subspace dimension is 12, while the counterpart of NPE is 7. The curves of recognition rate vs. dimension are shown in Fig.1. And the max recognition rates of each method are also reported in Tab.1.

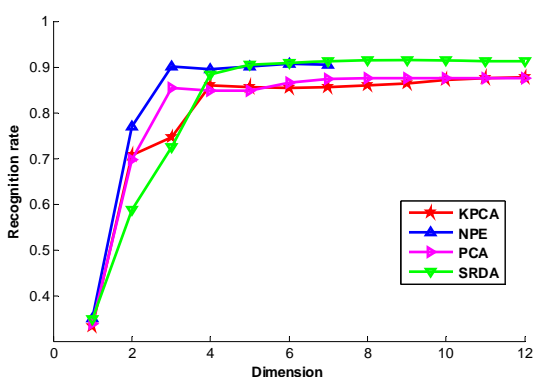


(a) 5 Training Samples





(c) 15 Training Samples



(d) 20 Training Samples

**Fig.1** The curves of recognition rate vs. Dimension of different training samples of each approaches

**Tab.1** The max recognition rates on Washington DC Mall(%)

Training Size	NPE	KPCA	PCA	SRDA
5×12	76.70	79.11	75.53	79.32
10×12	85.87	83.40	83.69	87.21
15×12	89.57	85.74	85.52	89.83
20×12	90.67	87.70	87.62	91.51

From Fig.1 and Tab.1, we can find that with increasing in the number of the training samples, the recognition rate increases. Also, we can notice that the SRDA shows the best performance around all the methods, that is to say, the SRDA, we proposed algorithm, always has the highest resolution.

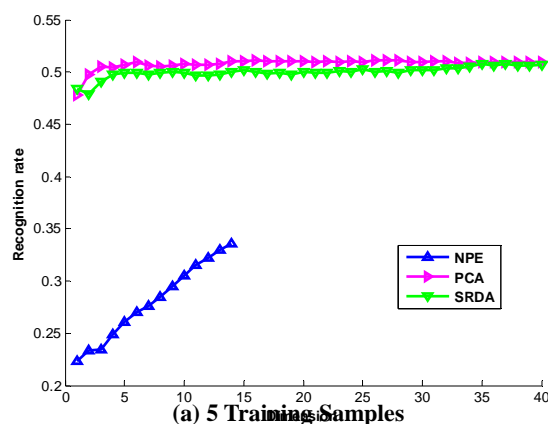
### 3.2 Experiments on AVIRIS Indian Pines

In this section, we conduct our experiments in following steps:

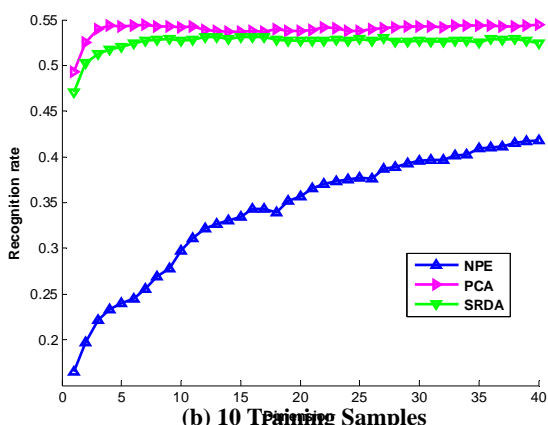
1) There are 16 labels in this dataset. In order to conduct the experiments reasonably, we reduce 6 labels whose sample numbers are all less than 320. Thus we need 320\*10 samples, and each sample contains 220 bands;

2) As described above, for the chosen 320 samples,  $l(=5, 10, 15, 20)$  datum are randomly selected for training and the rest 300 datum are used for testing;

3) The 5-nearest neighbor classifier is applied in PCA, NPE and SRDA subspace. For PCA and SRDA, the numbers of the subspace dimension are 40, while the NPE's are 14, 40, 40 and 40. The curves of recognition rate vs. dimension are shown in Fig.2. And the max recognition rates of each method are also reported in Tab.2.



(a) 5 Training Samples



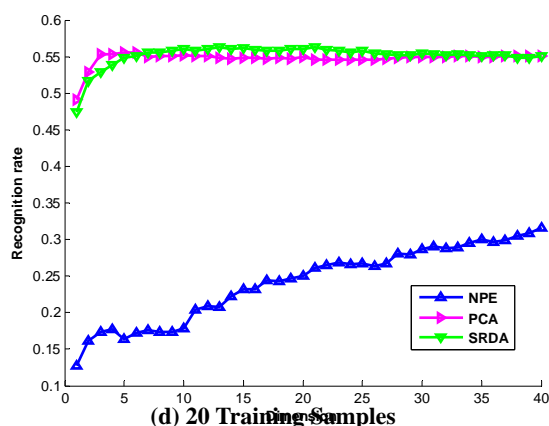
(b) 10 Training Samples

classification with the increase in the number of the training samples, expect NPE when it is applied in 5-nearest neighbor classifier.

2) The NPE, KPCA, and PCA are all involved with eigen-decomposition of dense matrices, which is computational expensive. While SRDA only needs to solve  $c-1$  regularized Least-Squares Problems which are efficient. Here the  $c$  represents the number of classes.

3) When the number of training samples is small, the same dimensions cannot be access to all these methods. So we have to change the dimensions to meet the demands of the experiments.

(c) 15 Training Samples



(d) 20 Training Samples

**Fig.2 The curves of recognition rate vs. Dimension of different training samples of each approaches**

**Tab.2 The max recognition rates on AVIRIS Indian Pines (%)**

Training Size	NPE	PCA	SRDA
5×220	33.55	51.16	50.79
10×220	41.79	54.47	53.14
15×220	36.96	56.02	55.27
20×220	31.54	56.22	56.31

From the Fig.2 and Tab.2, we also can observe that the SRDA also has a higher recognition rate, compared with the PCA and NPE, especially when the number of the training samples increases to 20, the recognition of SRDA is the highest. Also with the number of training samples turns large, the recognition rate increases, and this is true to PCA and SRDA, but not fit with the NPE. And this is related to the neighbor classifier.

### 3.3 Discussion

The experiments on Washington DC Mall and AVIRIS Indian Pines have reflected some significant points.

1) All methods mentioned in this paper shares higher

## 4. CONCLUSIONS

In this paper, we developed an efficient and useful approach for dimension reduction, which is called the Spectral Regression Discriminant Analysis (SRDA). This method avoids the difficulty of eigen-decomposition and casts the problem of learning an embedding function into a regression framework, which is a huge save of time and memory. As we all know that, the SRDA can conduct discriminant analysis of large-scale high-dimensional data. With experiments, we can easily find that the SRDA shows higher recognition rate by comparison to other methods such as NPE, LDA, PCA and KPCA.

## REFERENCES

- [1] R. Dianat, S. Kasaei, 2009. Dimension reduction of remote sensing images by incorporating spatial and spectral properties. *Int. J. Electron. Commun. (AEU)*, 64 (2010), pp. 729–732.
- [2] M. Belkin and P. Niyogi, 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing System 14*, pp. 585–591.
- [3] Jolliffe I. T., 2002. *Principal Component Analysis*, Springer, Berlin, pp. 150–166.
- [4] R. O. Duda, P. E. Hart, D. G. Stork, 2000. *Pattern Classification*, 2nd edition, Wiley-Interscience, NY, pp. 214–281.
- [5] M. Brand, 2003. Continuous nonlinear dimensionality reduction by kernel eigenmaps, In *International Joint Conference on Artificial Intelligence*.
- [6] X. F. He, D. Cai, S. C. Yan and H. J. Zhang, 2005. Neighborhood preserving embedding, In *Proceedings of IEEE 10th International Conference on Computer Vision*, Beijing, China, pp. 1208–1213.
- [7] Y. X. Li, X. G. Ruan, 2005. Cancer subtype recognition and feature selection with gene expression profiles. *Acta Electronica Sinica*, 33(4), pp. 651–655.
- [8] S. Roweis, L. Saul, 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), pp. 2323–2326.
- [9] J. Tenenbaum, V. de Silva, J. Langford, 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, pp. 2319–2323.

- [12] X. F. He, S. C. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, 2005. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3), pp. 328–340.
- [13] S. Yan, D. Xu, B. Zhang, and H. J. Zhang, 2005. Graph embedding: A general framework for dimensionality reduction. In Proc. CVPR 2005.
- [14] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), pp. 40–51.
- [15] D. Cai, X. F. He, and J. W. Han, 2007. Spectral Regression for Dimensionality Reduction. *Department of Computer Science, Technical Report NO.2865*, University of Illinois at Urbana-Champaign.
- [16] G. H. Golub and C. F. V. Loan, 1996. *Matrix computations*. Johns Hopkins University Press, 3rd edition.
- [17] T. Hastie, R. Tibshirani, J. Friedman, 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.
- [18] D. Cai, X. F. He, and J. W. Han, 2007. Spectral regression for dimensionality reduction. *Technical report Computer Science Department, UIUC, UIUCDCS-R-2007-2856*.
- [19] D. Cai, X. F. He, J. W. Han, 2008. SRDA: An efficient algorithm for large scale discriminant analysis. *IEEE Transactions on knowledge and data engineering*, 20(1), pp. 1-12.
- [20] F. R. K. Chung, 1997. *Spectral Graph Theory*, volume 92 of *Regional Conference Series in Mathematics*, AMS.